# Lab_5_GroupReport

*Akshayaswuroupikka Balasubramanian,yixuan xu*

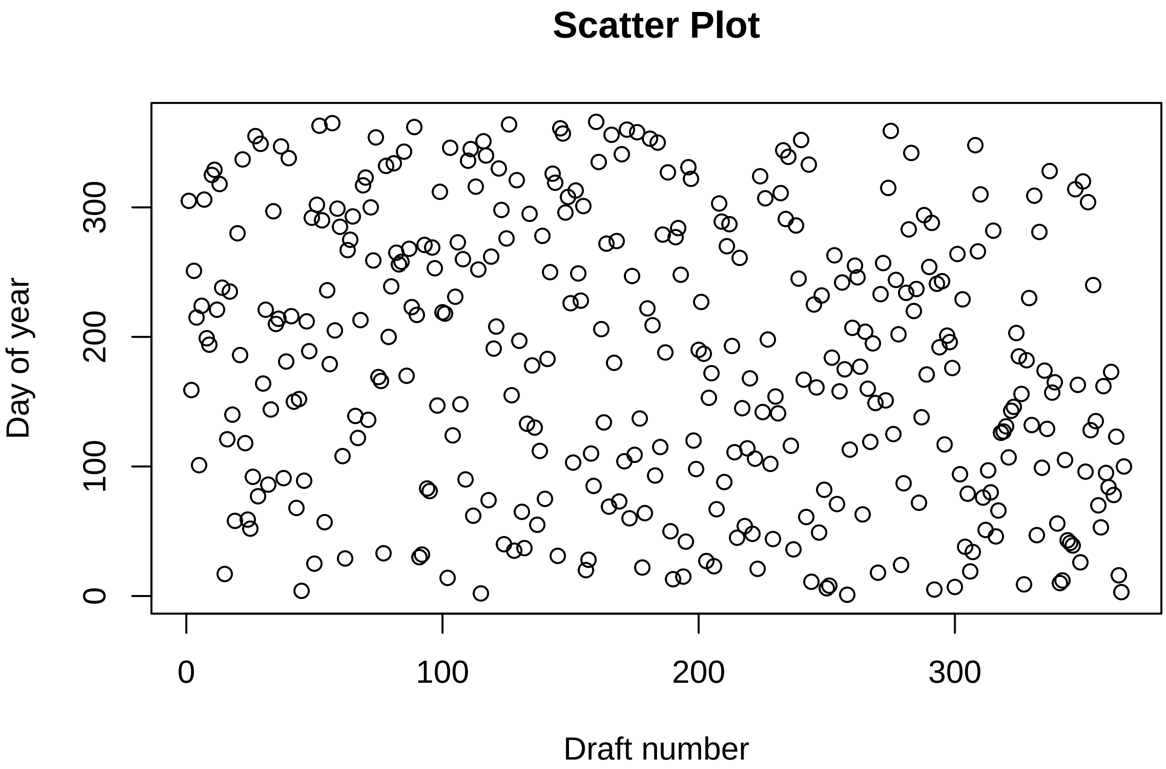
*March 9, 2016*

**Assignment 1: Hypothesis testing**

## In 1970, the US Congress instituted a random selection process for the military draft. All 366 possible birth dates were placed in plastic capsules in a rotating drum and were selected one by one. The first date drawn from the drum received draft number one, the second date drawn received draft number two, etc. Then, eligible men were drafted in the order given by the draft number of their birth date. In a truly random lottery there should be no relationship between the date and the draft number. Your task is to investigate whether or not the draft numbers were randomly selected. The draft numbers (Y=Draft_No) sorted by day of year (X=Day_of_year) are given in the file lottery.xls

## 1.1Make a scatterplot of Y versus X and conclude whether the lottery looks random.

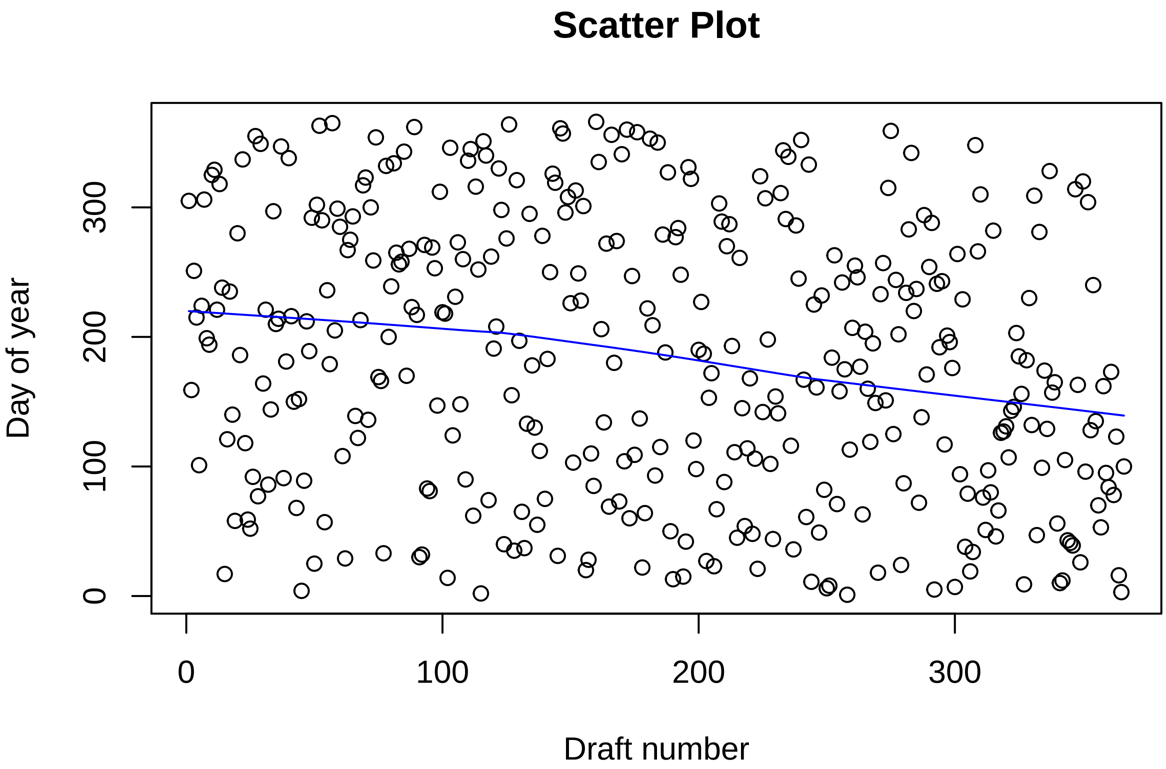
The scatterplot of Y(Draft_no) verses X(Day_of_year) is plotted below:



From the scatterplot one could say that the lottery looks random.There is no trace of a trend or a pattern which can be observed.The values are spreadout or scattered all through the graph.

## 1.2 Compute an estimate $\hat{y}$???????? of the expected response as a function of X by using a loess smoother (use loess() ), put the curve $\hat{y}$???????? versus X in the previous graph and state again whether the lottery looks random.

An estimate of the expected response as a function of X by using a loess smoother (use loess() )is computed below and the curve versus X is plotted in the previous graph.



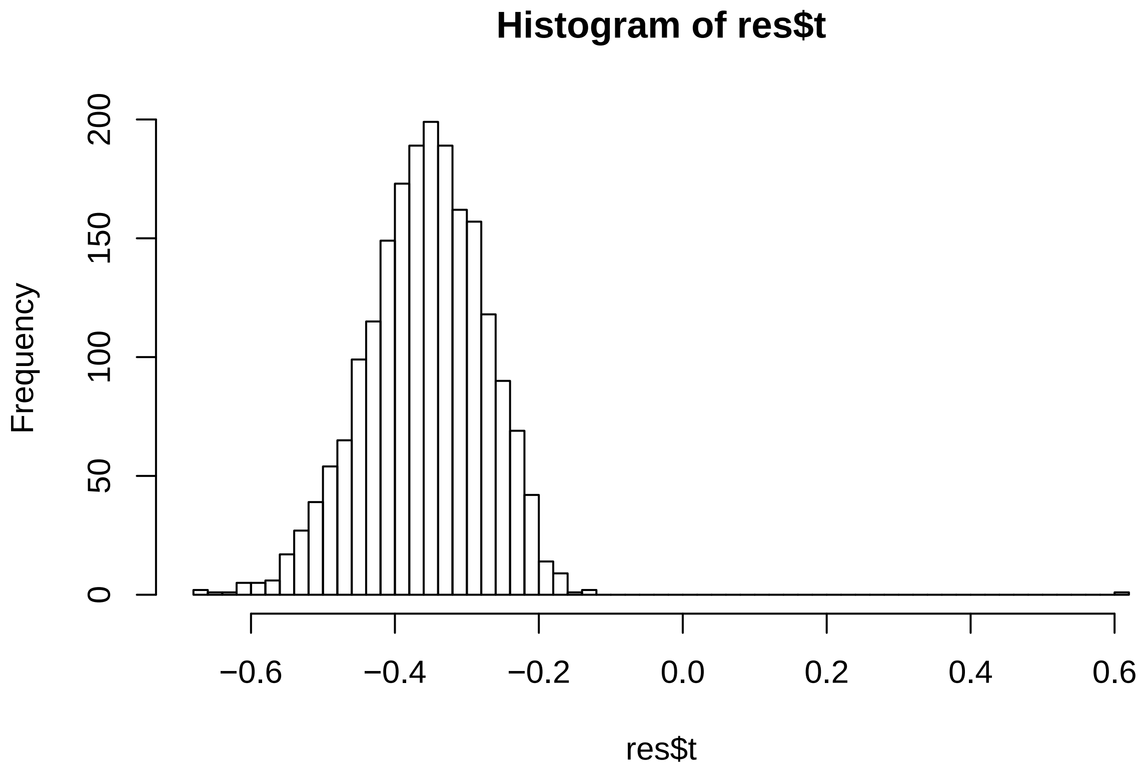
There seem to be a negative trend.This trend causes uncertainty over the randomness for the lottery procedure.

## 1.3To check whether the lottery is random, it is reasonable to use test statistics

$T = \frac{\hat{Y}(X_b)-\hat{Y}(X_a)}{X_b-X_a}, Where X_b = argmax_X \hat{Y}, X_a = argmin_X \hat{Y}$

## If this value is significantly greater than zero, then there should be a trend in the data and the lottery is not random. Estimate the distribution of ???? by using a non-parametric bootstrap with ????=2000 and comment whether the lottery is random or not. What is the P-value of the test?

To check whether the lottery is random, it is reasonable to use test statistics $T = \frac{\hat{Y}(X_b)-\hat{Y}(X_a)}{X_b-X_a}, Where X_b = argmax_X \hat{Y}, X_a = argmin_X \hat{Y}$ The distribution of T by using a non-parametric bootstrap with B=2000 is estimated.

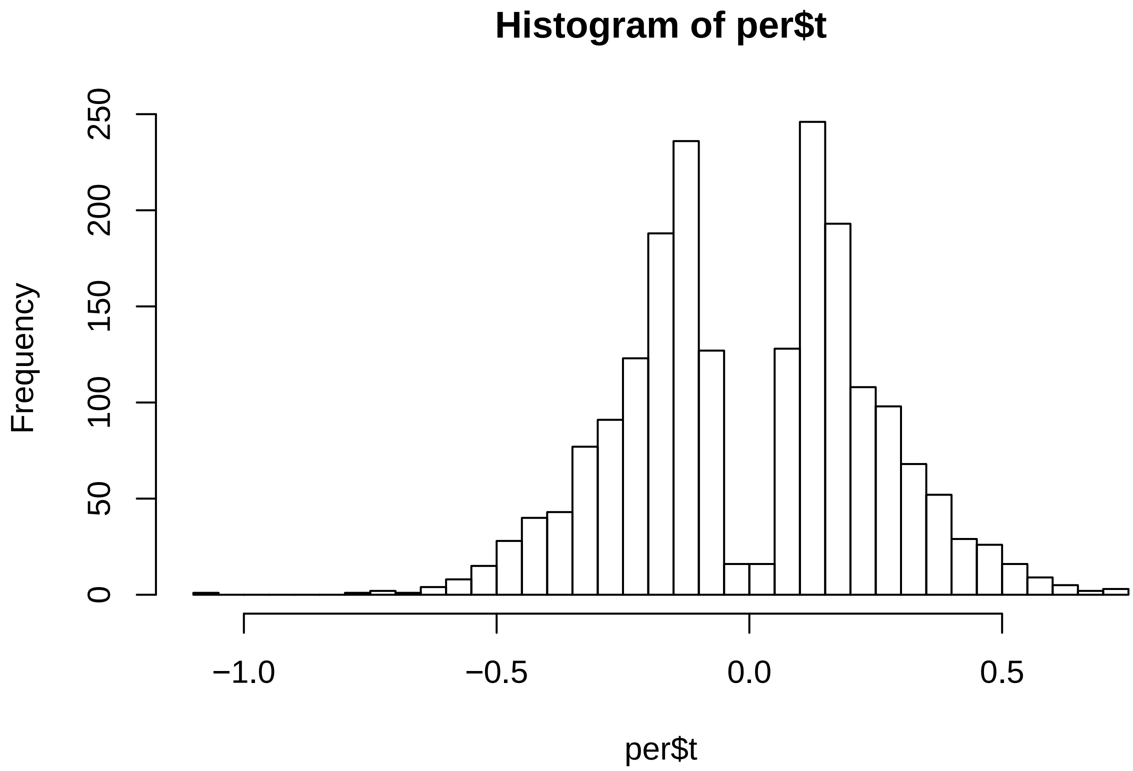


```
## [1] 0.6092419
```

The corresponding histogram is plotted and a T value greater than Zero signifies that there is a trend in data which means the lottery is not random.The estimated P-value is r' p_value'.

## 1.4 1.4 Implement a function depending on data and B and that tests the hypothesis $H_0$ *Lottery is random* vs $H_a$ *Lottery is not random* by using a permutation test with statistics T and returns the p-value of this test. Test this function on our data with B=2000.

A function depending on data and B and that tests the hypothesis by using a permutation test with statistics is written below:



This function is tested with B=2000.The returned P-value is r' permutation_test(lottery, 2000)$p_value' .Depending on the choice of significance level the conclusion about the randomness differ since the P-values is relatively low. However, if a the significance level is chosen to be 0.05, a common choice, the conclusion is that $H_0$ cannot be rejected (the lottery is random).

## 1.5Make a crude estimate of the power of the test constructed in step 4:

**a. Generate (an obviously non-random) dataset with $n = 366$ observations by using same X as in the original data set and $Y(x) = max(0, min(\alpha x + \beta,\ 366))$ where $\alpha = 0.1$ and $beta \sim N(183, sd = 10)$**

**b. Plug these data into the permutation test with B=200 and note whether it was rejected**

**c. Repeat steps a)-b) for $\alpha = 0.2, 0.3, \cdots, 10$**

**What can you say about the quality of your test statistics considering the value of the power?**

```
## [1] 0.005
```

```
## [1] "reject"
```

```
##  [1] "reject" "reject" "reject" "reject" "reject" "reject" "reject"
##  [8] "reject" "reject" "reject" "reject" "reject" "reject" "reject"
## [15] "reject" "reject" "reject" "reject" "reject" "reject" "reject"
## [22] "reject" "reject" "reject" "reject" "reject" "reject" "reject"
## [29] "reject" "reject" "reject" "reject" "reject" "reject" "reject"
## [36] "reject" "reject" "reject" "reject" "reject" "reject" "reject"
## [43] "reject" "reject" "reject" "reject" "reject" "reject" "reject"
## [50] "reject" "reject" "reject" "reject" "reject" "reject" "reject"
## [57] "reject" "reject" "reject" "reject" "reject" "reject" "reject"
## [64] "reject" "reject" "reject" "reject" "reject" "reject" "reject"
## [71] "reject" "reject" "reject" "reject" "reject" "reject" "reject"
## [78] "reject" "reject" "reject" "reject" "reject" "reject" "reject"
## [85] "reject" "reject" "reject" "reject" "reject" "reject" "reject"
## [92] "reject" "reject" "reject" "reject" "reject" "reject" "reject"
## [99] "reject"
```

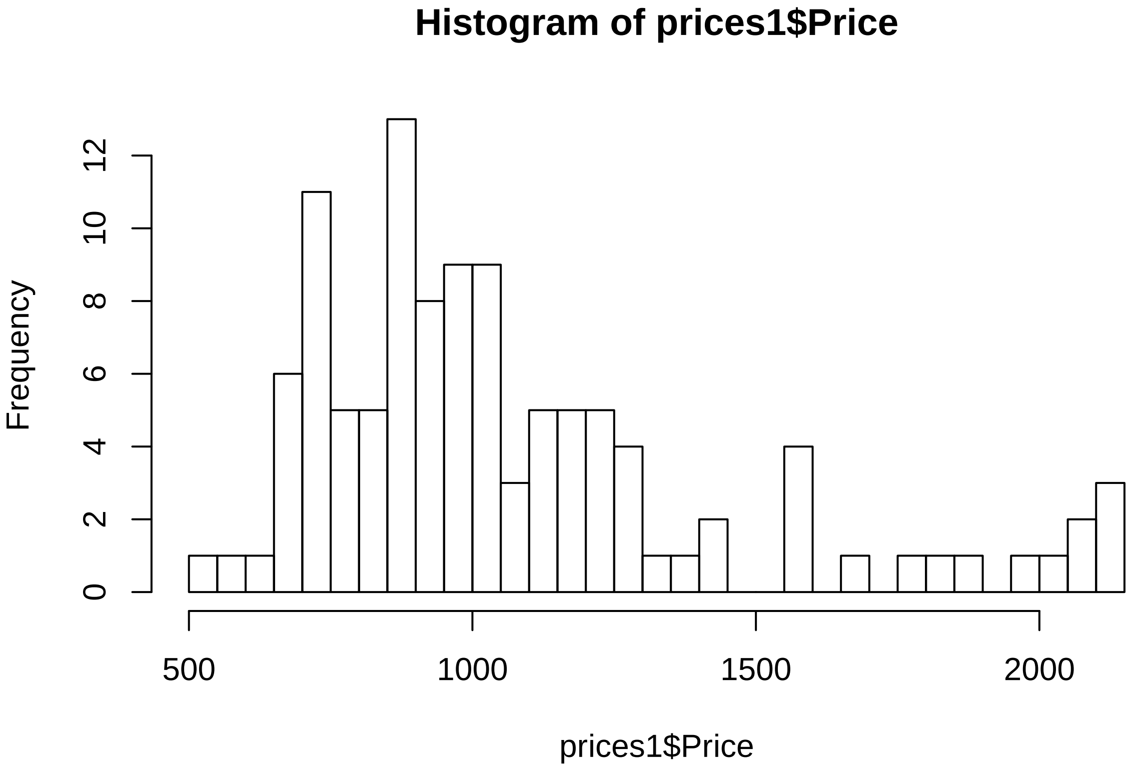
As the values which are given are rejected, the quality of the test statistics is 100 %.

# Assignment 2: Bootstrap, jackknife and confidence intervals

**The data you are going to continue analyzing is the database of home prices in Albuquerque, 1993.**

**The variables present are Price; SqFt - the area of a house; FEATS - number of features such as dishwasher, refrigerator and so on; Taxes - annual taxes paid for the house. Explore the file prices1.xls**

**2.1 Plot the histogram of Price. Does it remind any conventional distribution? Compute the mean price.**



```
## [1] 1080.473
```

The conventional distribution looks mostly like chi-square distribution or gamma distribution.(chi-square distribution is a special of gamma distribution.)

### 2.2 Estimate the distribution of the mean price of the house using bootstrap. Determine the bootstrap bias-correction and the variance of the mean price. Compute the 95% confidence interval for the mean price using bootstrap percentile, bootstrap BCa, and first-order normal approximation (Hint: use *boot()*,*boot.ci()*,*plot.boot()*,*print.bootci()*)

Since I am not sure about distribution of house price, so non-parametric bootstrap could be used in this situation.

```
## [1] "bootstrap bias-correction:" "1079.60856363636"
```

```
## [1] "variance of the mean price:" "1372.51893299013"
```

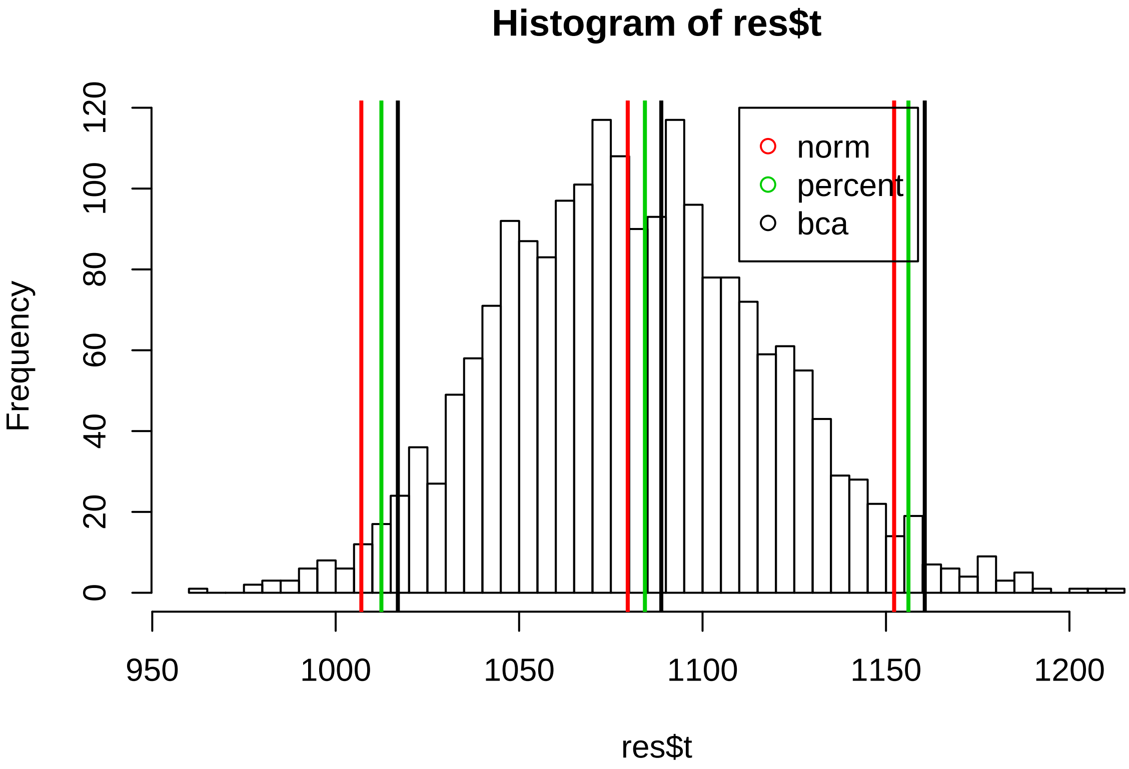
```
##            lower    upper     mean   length
## norm    1006.997 1152.220 1079.609 145.2236
## percent 1012.460 1156.121 1084.290 143.6610
## bca     1016.951 1160.568 1088.759 143.6168
```

### 2.3 Estimate the variance of the mean price using the jackknife and compare it with the bootstrap estimate

```
## [1] 1320.911
```

The variance using jackknife is larger than using bootstrap, that is normal happened, cause the variance using jackknife is often overestimated.

### 2.4 Compare the confidence intervals obtained with respect to their length and the location of the estimated mean in these intervals.



The confidence intervals matrix was shown in the step 2, and the mean value from first-order normal is close to the real mean value, and other method is far away from the real value, BCa value is larger than percent value. The length of each method, BCa has the largest value, and normal and percent is very close to each other. Method BCa and normal intervals looks synmetric, but the percentile intervals is not synmetric.

## contribution

The first part of the assignment is contributed by akshaya and the second part by Yixuan xu.

# Appendix - R-code

```r
## ----echo=FALSE, message=FALSE-------------------------------------------
require(XLConnect)
library(boot)
data= loadWorkbook("lottery.xls")
lottery = readWorksheet(data,sheet = "sheet1", header = TRUE)
plot(lottery$Draft_No~lottery$Day_of_year,main="Scatter Plot",
xlab="Draft number",ylab="Day of year")
```

```
## ---- echo=FALSE---------------------------------------------------------
plot(lottery$Draft_No~lottery$Day_of_year,main="Scatter Plot",
xlab="Draft number",ylab="Day of year")
lines(lowess(lottery$Draft_No,lottery$Day_of_year), col="blue")

## ---- echo=FALSE---------------------------------------------------------
test_statistics<-function(data,n){
  data=data[n,]
  lm=loess(Draft_No ~ Day_of_year, data=data)
  Xa=which.min(lm$fitted)
  Xb=which.max(lm$fitted)
  y_Xa=lm$fitted[Xa]
  y_Xb=lm$fitted[Xb]
  ts= (y_Xb - y_Xa)/(data$Day_of_year[Xb] - data$Day_of_year[Xa])
  return(ts)
}
set.seed(311015)
res=boot(lottery,test_statistics,R=2000)
hist(res$t,50)
for( i in 1:length(res$t)){
  if(res$t[i] >0){
    print(res$t[i])
  }
}

p_value=mean(res$t>0)

## ----echo=FALSE----------------------------------------------------------
permutation_test<- function(data,B){
  ts=numeric(B)
  n=dim(data)[1]
  for(i in 1:B){
    doy=sample(data$Day_of_year, n, replace = FALSE)
    new_data=data.frame(Draft_No=data$Draft_No,Day_of_year=doy)
    lm=loess(Draft_No ~ Day_of_year,new_data)
    Xa=which.min(lm$fitted)
    Xb=which.max(lm$fitted)
    y_Xa=lm$fitted[Xa]
    y_Xb=lm$fitted[Xb]
    ts[i]=(y_Xb - y_Xa)/(new_data$Day_of_year[Xb] - new_data$Day_of_year[Xa])
  }
  lm1=loess(Draft_No ~ Day_of_year, data=data)
  Xa1=which.min(lm1$fitted)
  Xb1=which.max(lm1$fitted)
  y_Xa1=lm1$fitted[Xa1]
  y_Xb1=lm1$fitted[Xb1]
  ts1=(y_Xb1 - y_Xa1)/(Xb1 - Xa1)
  p_value<- mean(abs(ts) > abs(ts1))
  return(list(p_value=p_value,t=ts))
}
set.seed(311015)
per=permutation_test(lottery,2000)
hist(per$t,50)
```

```
## ----echo=FALSE----------------------------------------------------------
lottery$Draft_No=0
for(i in 1:366){
  lottery$Draft_No[i] <- max(0, min((0.1*lottery$Day_of_year[i] + rnorm(1, 183, 10)), 366))
}
permutation_test(lottery, 200)$p_value

alpha=seq(0.2, 10, 0.1)
p_value=0
for(i in 1:length(alpha)){
  lottery$Draft_No=0
  for(j in 1:366){
    lottery$Draft_No[j]= max(0, min(alpha[i]*lottery$Day_of_year[j] + rnorm(1, 183, 10), 366))
  }
  p_value[i] <- permutation_test(lottery, 200)$p_value
}


ifelse(p_value<=0.05,print("reject"),print("not reject"))


## ----echo=FALSE----------------------------------------------------------
require(XLConnect)
data1= loadWorkbook("prices1.xls")
prices1=readWorksheet(data1, sheet = "sheet1", header = TRUE)
hist(prices1$Price,50)
mean(prices1$Price)

## ---- echo=FALSE---------------------------------------------------------
library(boot)
f <- function(data,n){
  data1 =data[n,]
  res <- mean(data1$Price)
  return(res)
}
res <- boot(prices1,f,R=2000)
print(c("bootstrap bias-correction:",2*mean(prices1$Price)-mean(res$t)))
# sum((res$t-mean(res$t))^2)/(length(res$t)-1)
print(c("variance of the mean price:",var(res$t)))
ci <- boot.ci(res,type=c("norm","perc", "bca"))
interval <- data.frame(rbind(ci$norm[2:3],ci$percent[4:5],ci$bca[4:5]),c(mean(ci$norm[2:3]), mean(ci$pe
rownames(interval)<- c("norm","percent","bca")
colnames(interval)<- c("lower","upper","mean","length")
interval

## ---- echo=FALSE---------------------------------------------------------
r <- length(prices1$Price)
jstar <- NULL
for(i in 1 : r){
jstar[i] <- r*mean(prices1$Price)-(r-1)*mean(prices1$Price[-i])
}
jt <- mean(jstar)
vart <- sum((jstar-jt)^2)/(r*(r-1))
```

```
vart

## ---- echo=FALSE---------------------------------------------------------

hist(res$t, 50)
abline(v = interval$lower, col = as.factor(rownames(interval)),lwd = 2)
abline(v = interval$upper, col = as.factor(rownames(interval)),lwd = 2)
abline(v = interval$mean, col = as.factor(rownames(interval)),lwd = 2)
legend(x=1110, y = 120,rownames(interval),col = as.factor(rownames(interval)),pch = 1)

## ----code=readLines(knitr::purl("lab_5_GroupReport.Rmd", documentation = 1)), eval = FALSE----
## NA
```